 MOTION IMAGERY STANDARDS BOARD	MISB RP 0904.4
RECOMMENDED PRACTICE	
Bandwidth-Quality-Latency Tradeoffs for Compressed Motion Imagery	1 November 2018

1 Scope

As high definition (HD) sensors become more widely deployed in the infrastructure, the migration to HD is straining communications channels and networks. Rather than accept either less fidelity in received images from over compression, or significantly increasing the bandwidth of these networks, this Recommended Practice offers guidance on methods to leverage HD Motion Imagery regardless of the limits in delivery. These methods include: cropping, scaling, frame decimation and compression coding structure.

This document addresses tradeoffs in image quality and latency of compressed Motion Imagery when constrained in channel bandwidth. The guidelines are based on subjective evaluations using an industry software encoder and several commercial hardware encoders. Data compression is highly dependent on scene content complexity, and for this reason the evaluation is based on two types of content: 1) panning over a multiplicity of high contrast, fast moving objects (people) and fine-detailed buildings; and 2) aerial imagery of planes, ground vehicles, and terrain typical of UAS collects. While the derived data rates may not reflect all types of scene content, they do serve as practical baselines. Vendors are encouraged to validate the practical implementation of the processing methods suggested.

Note on image nomenclature: Image formats discussed include progressive-scan imagery only. For this reason, the “p” generally applied as a suffix when describing progressive-scan formats (for example, 1080p and 720p) is suppressed.

2 References

- [1] MISB MISP-2019.1 Motion Imagery Standards Profile, Nov 2018.
- [2] MISB RP 0802.2 H.264/AVC Motion Imagery Coding, Feb 2014.
- [3] MISB RP 1011.1 LVSD Motion Imagery Streaming, Feb 2014.

3 Acronyms

FOV	Field of View
FPS	Frames per second
GOP	Group of Pictures
HD	High Definition
TCDL	Tactical Common Data Link

SD Standard Definition

4 Revision History

Revision	Date	Summary of Changes
RP 0904.4	11/1/2018	<ul style="list-style-type: none"> Generalized to include other compression types Title change Remove reference to MISP POI table data rates and quality as this information is no longer supplied Updated MISP Reference [1]

5 Introduction

Consider an adjustable, motion imagery encoder of Figure 1 designed to accommodate a prescribed data link bandwidth.

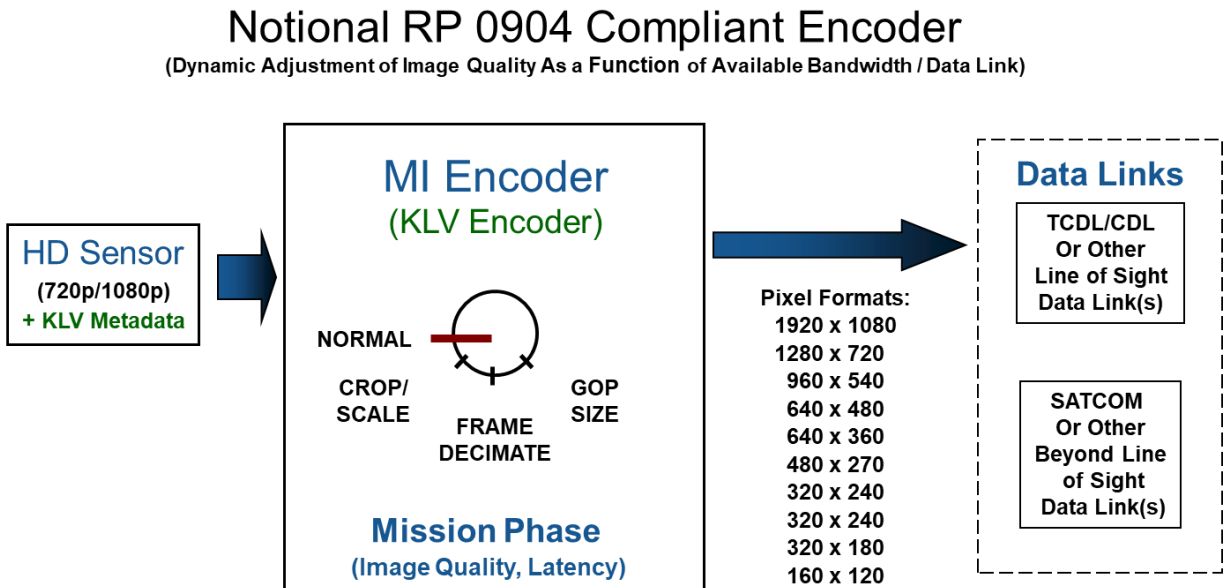


Figure 1: Adjustable Motion Imagery Format HD Encoder

Here, a high definition sensor produces High Definition (HD) content of 1920x1080 or 1280x720 format. An operator can set the encoder to meet a specific channel data rate (if the channel rate is known), or the encoder can be set automatically if data link capacity is actively sensed and fed back. In either case, there are numerous options to choose when compressing the HD source motion imagery. In the **NORMAL** mode, the encoder compresses the HD sensor data as it is received. In cases where the channel bandwidth will not support the encoding of HD to sufficient image quality, pre-processing the content by cropping and/or scaling (**CROP/SCALE**) and/or decimating frames (**FRAME DECIMATE**) may provide the necessary image quality for the CONOPS needed. An additional option is to choose an encoding GOP (Group of Pictures) size

(**GOP SIZE**) which reduces the compressed data rate. Numerous spatial formats have been selected to maximize interoperability.

Figure 1 suggests a structure for altering the image format and mission requirements as a function of available data link bandwidth. Beyond meeting the data link requirements this new functionality provides versatility in changing the image quality based on real-time in-flight mission needs.

Table 1 lists some of the effects in applying these different capabilities to meet a given channel bandwidth. It is assumed the channel bandwidth does not support sufficient quality imagery from a sensors native image format; that is, the image content is over-compressed filled with compression artifacts. One or more of the capabilities listed may be applied to the imagery; crop, scale and frame decimation. These are performed on the image sequence prior to compression, while setting a longer GOP (Group of Pictures) is an internal encoder parameter, and is effective during the compression process.

Table 1: Capabilities and Effects on Compressed Stream

Conditions	
<ul style="list-style-type: none"> • Channel bandwidth fixed • Native image format compressed; image quality very poor 	
Capability	Effect
Cropping	<ul style="list-style-type: none"> • Reduced Field of View • Image quality improved • Latency the same (assumes processing time insignificant)
Scaling	<ul style="list-style-type: none"> • Equivalent Field of View • Image quality improves; however, reduced spatial frequencies (potentially less detail than original) • Latency the same (assumes processing time insignificant)
Frame Decimation	<ul style="list-style-type: none"> • Image quality improves (assuming low motion content) • Latency increases (time between frames increases)
Longer GOP Length	<ul style="list-style-type: none"> • Image quality improves • Latency reduced • Susceptible to transmission errors without Intra-refresh • Longer start up on some decoders (waiting for I-Frame)

6 High Definition (HD) Format

The High Definition (HD) spatial formats are 1920 horizontally by 1080 vertically, and 1280 horizontally by 720 vertically; these have a maximum frame rate of 60 frames-per-second (FPS). Both formats have square pixels (PAR = 1:1), which means the ratio of horizontal to vertical size of each pixel is 1:1. When given a high definition sensor source, a compressed high definition, high quality image can be delivered when there is enough channel bandwidth. However, in cases where the channel bandwidth is insufficient, over-compressing the HD sequence will only produce severely degraded images. Several capabilities to reduce the data rate and improve the

image quality are listed in Table 1. These approaches do impact an encoders design, and not every option may be available from a given manufacturer.

7 Cropping and Scaling

7.1 Aspect Ratio

To better appreciate the consequences of cropping and scaling it is best to review terminology used in the industry. There are three different associations for the words “aspect ratio” as found in the literature. The *Pixel Aspect Ratio (PAR)* is expressed as a fraction of the horizontal (x) pixel size divided by the vertical (y) pixel size. For example, the PAR for square pixels is 1:1.

The *Source Aspect Ratio (SAR)* is the ratio of total horizontal (x) picture size to total vertical (y) image size, for a stated definition of “image.” SAR can be equated to the format of what the sensor or source of content is. For example, a high-definition sensor has a SAR of 16:9.

Finally, there is the *Display Aspect Ratio (DAR)*. The DAR refers to the display or monitor aspect ratio of width to height. Appendix A provides some examples of these three aspect ratio metrics and how they relate to one another. While these are all factors to consider, perhaps the most relevant to cropping and scaling is the pixel aspect ratio PAR.

Cropping any arbitrary region will always produce an image with the same PAR as the source image. For example, a 640x480 image cropped from a high definition image will have square pixels. Scaling, on the other hand, will affect the PAR if the scaling is not done equally in both the horizontal and vertical dimensions. For example, a 640x360 image scaled down from a 1280x720 image will have the same PAR (1:1) (scaled by 1/2 in each dimension); a 640x480 image scaled from the same 1280x720 image will have a PAR of 1:0.75.

7.2 Image Cropping

Cropping preserves the pixel aspect ratio of the source image. So, if a 4:3 original image has non-square pixels, then a cropped sub-image will also have non-square pixels. Likewise, if a 16:9 image has square pixels, then a cropped sub-image will also have square pixels. In image cropping, a smaller sub-area *within* the sensors field of view is extracted for encoding. For example, as shown in Figure 2, if the HD sensor field of view is 1280x720 (horizontal pixels x vertical pixels), extracting a sub-area of 640x480 produces imagery with equivalent pixels to the original imagery within the respective sub-area. This reduced-size image represents a reduced field of view with respect to the original. In this case, the 1:1 pixel-aspect ratio of the HD source image is maintained, so geometric distortion will not occur – i.e. circles in the original remain circles in the sub-area image. As indicated in Figure 2 source image content outside the cropped area is lost.

It is to be cautioned that cropping affects metadata which may describe the source image characteristics; particularly, image coordinates and other positional data and information regarding the geometry of pixels. When cropping, additional metadata should indicate cropping has occurred, and the source metadata needs to be corrected. Knowledge of the original and resulting image size would allow metadata, such as corner points, to be recalculated at the receiver.

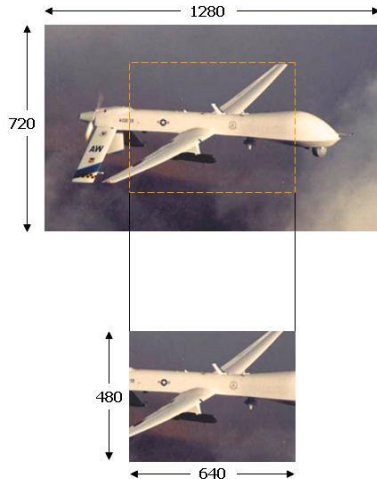


Figure 2: Image Cropping Example

Sub-image extracted from the full image field, where the pixels within the sub-image are equivalent to those within the original image.

7.3 Image Scaling

In image scaling, the sensors field of view (FOV) is preserved, but possibly at the expense of the spatial frequency content, which likely will be reduced. This may result in a loss of fine image detail. For example, Figure 3 shows an HD sensor field of view with a format of 1920x1080 pixels scaled by one-half in each dimension to produce an image with a format of 960x540 pixels. Note the output image looks identical to the input, except smaller. To preserve the SAR (horizontal to vertical size) each image dimension is scaled by the equivalent amount. This will ensure geometric shapes like circles in the original image remain circles in the scaled image. Square pixels are preserved.

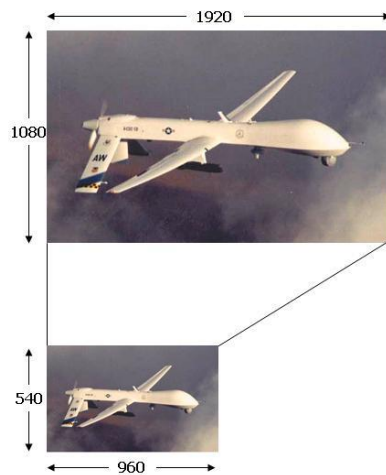


Figure 3: Image Scaling Example

New image filtered and scaled, where the original field of view is maintained.

While image cropping requires nothing more than a simple remapping of input pixels to those within a target output sub-area, scaling requires spatial pre-filtering of the image. Simple techniques such as pixel decimation and bilinear filtering can produce image artifacts: in pixel decimation image-aliasing can cause false image structure, which also impacts the compression negatively; bilinear filtering may produce excessive blurring, particularly for large scale factors. More information on filter guidelines can be found in Appendix B.

7.4 Image Crop & Scale – HD to SD

Illustrated next in Figure 4 is an example of combining cropping and scaling to convert a 1920x1080 HD image to a 640x480 SD image. The goal is to maintain the square pixel relationship of the original image in the scaled image, so there is no geometric distortion. To do so necessitates a certain amount of the original image be cut off; this can be done equally to each side as shown in the example, or taken completely from one side or the other, thereby skewing the image to that side.

This type of conversion is very typical of current home experiences in watching high definition content on a standard definition television receiver. The image on each side is cut off and not visible to those with standard definition receivers.

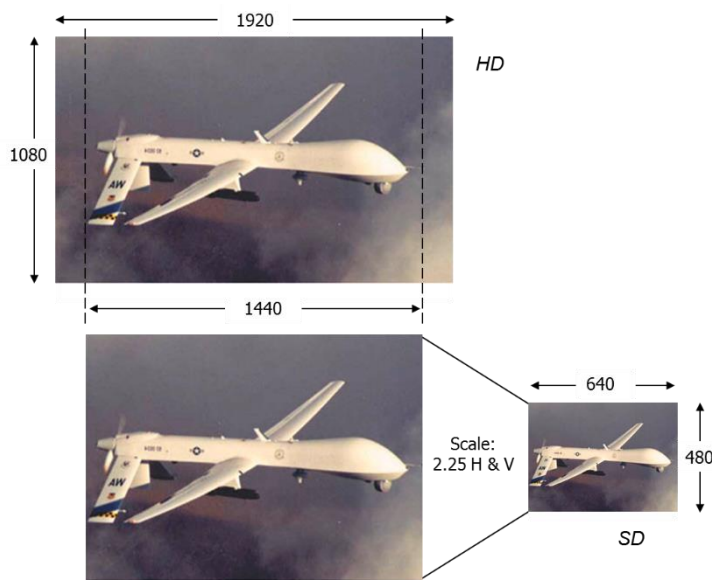


Figure 4: Image Crop & Scale Example

1920x1080 HD image is first cropped to 1440x1080, and then equally scaled by 4/9 both horizontally and vertically to produce a 640x480 pixel SD format image suitable for display on a 4:3 display.

7.5 Frame Rate Decimation

Another option for reducing the data rate is to eliminate image frames; this is called frame decimation. Dropping every other frame will produce a source sequence of one-half the original frame rate; for example, 30 frames-per-second (FPS) to 15 FPS. Dropping two frames out of every three of a 30 FPS sequence will produce a 10 FPS sequence. Removing frames should be done carefully. When the content has a high degree of motion removing frames may cause temporal aliasing, which produces artifacts on image edges of the moving objects. In the absence of high motion, dropping frames will allow the encoder to spend its bits on the remaining images; this should improve image quality.

One issue with removing frames is the distance in time between the frames increases. For example, at 30 FPS the time between frames is 1/30 second; at 15 FPS, the time between frames is 1/15 second. This causes latency in processing. In general, higher frame rates demand more bits to encode (there is more data to compress), but the latency is lower; whereas for lower frame rates more bits can be spent on the existing frames thereby increasing image quality, but latency

is increased. Finally, the impact to the source metadata must be considered when discarding frames. When frames are discarded, for example, changing the temporal rate from 60 to 30 frames per second, metadata associated with the dropped frames may also be dropped.

8 Objective Spatial Formats

While nearly any crop or scale can be applied to source imagery, the MISB has selected several spatial/temporal formats, which when used, provide for maximal interoperability. These specific formats, called Points of Interoperability (see MISB [1]), are encouraged in meeting desired spatial/temporal image formats.

The formats are independent of the choice to crop or scale. For example, if a source is 1280x720 pixels, this can be cropped to either a 640x320 or scaled to 640x320 image. In the cropped case, only a portion of the original field of view survives but may provide better detail fidelity in the resulting compressed image. In the scaled case, the entire field of view is preserved; it will be a smaller version of the original with a potential reduction in fine detail (less fidelity).

9 Longer GOP Size

GOP (Group of Pictures) is a mixture of I, B and P frames to form a repeated coding structure in MPEG compression. B-frames are typically not used when low latency is desired, so the discussion here is limited to I and P frame coding only. A GOP starts with an I-frame (intra-frame coded image) and includes all successive P-frames (Predicted) up to, but not including, the next I-frame. For a GOP size of 25 there would be a sequence of one I-frame followed by 24 P-frames. This pattern would then repeat throughout the remaining image sequence.

I-frames are expensive bit-wise; they require the most data to represent them. Thus, the fewer I-frames in the stream the less data produced in the compressed output stream. Viewed differently, for a given data rate the encoder can expend more bits encoding P-frames when there are fewer I-Frames, which results in a higher level of image quality. Making a longer GOP size suggests for a given coded sequence the overall data rate will be lower; or better image quality is possible.

Since I-frames are much larger than P-Frames, buffering is required in the encoder and the decoder to achieve a constant bit rate and to prevent decoder underrun. Larger GOPS typically will reduce this difference, thus reducing the buffering needed and reducing the associated latency. The limit of this is Infinite GOP (all P-Frames), which requires the least buffering and thus has the minimum latency.

The downside? Long GOP sequences are more prone to transmission errors. Because an I-frame is self-contained (no dependence of pre-or post-frames) their presence assures errors in a stream terminate and are corrected at the I-frame (assuming the errors are not in an I-frame). Intra-refresh is a coding tool intended to more quickly repair errors in a stream; this is a topic discussed in greater detail MISB RP 0802 [2] and MISB RP 1011 [3]. Another issue with long GOP sequences is it takes longer to start the decoding of a sequence, since decoding can only begin with an I-frame. This additional delay is only experienced upon tuning into a stream, and does not affect subsequent decoding latency.

10 Conclusions

The guidelines presented here offer suggested image formats and options based on current knowledge of product capabilities and performance. As the assumptions made here become tested, this document will refine its guidelines accordingly. Users need to make tradeoffs between image quality and latency. Reducing latency generally results in a lower image quality for a given data rate. When the lowest possible latency is required for a given bandwidth image spatial format reductions may be necessary.

Appendix A: Aspect Ratio Types - Informative

This appendix provides some examples to better appreciate aspect ratio and how it applies to the source, display and pixel geometry of imagery. Figure 5 defines two types of “aspect ratio”: Source Aspect Ratio (SAR) and Pixel Aspect Ratio (PAR). The Source could be the sensors native image spatial format.

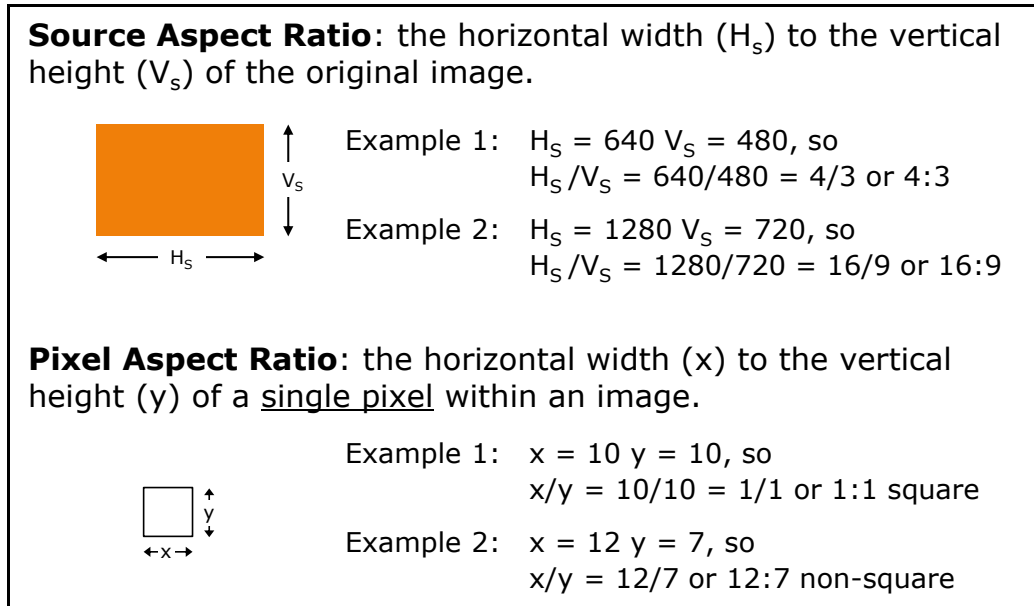


Figure 5: Definitions for SAR and PAR

In Figure 6, a third type of aspect ratio is defined: Display Aspect Ratio (DAR), which describes the aspect ratio of the display device, such as 4:3 for a NTSC display or 16:9 for an HD display.

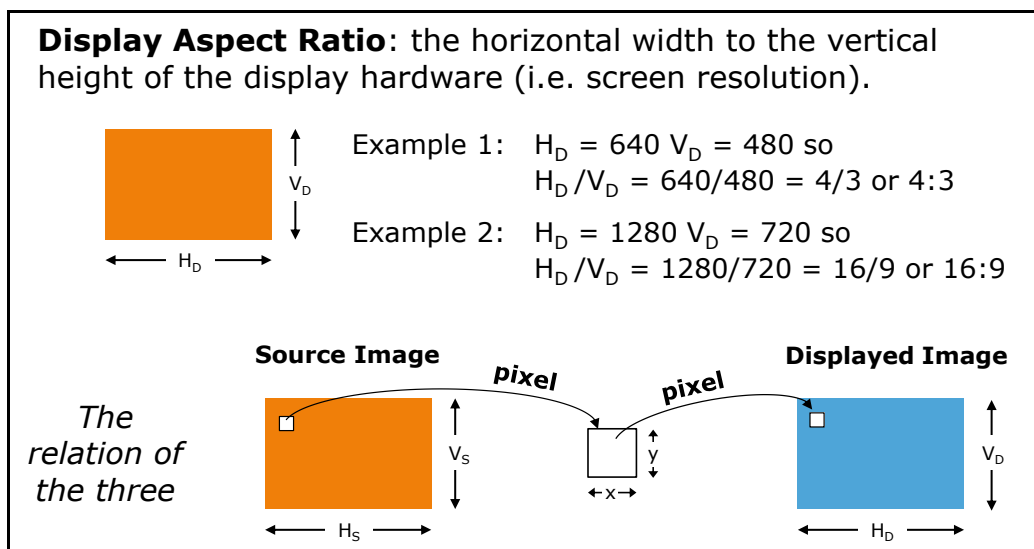


Figure 6: Definition of DAR

In Figure 7 the relationship among the three aspect ratio types show multiplying the SAR with PAR yields the Display Aspect Ratio, which provides a measure of distortion.

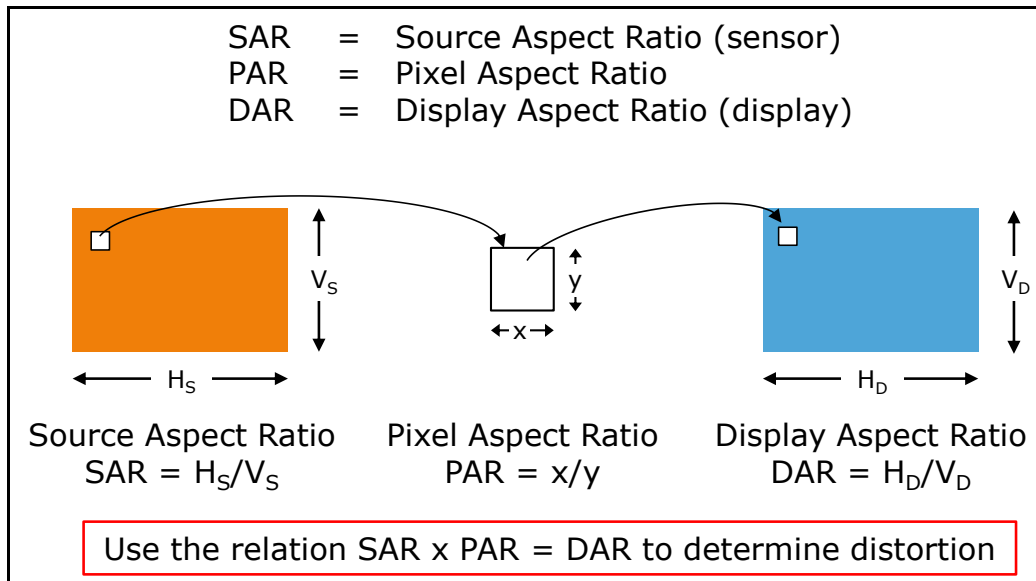


Figure 7: Relationship among SAR, PAR and DAR

Figure 8 presents an ideal case where the SAR and DAR are the same so the $PAR = 1:1$. This results in a one-to-one pixel mapping requiring no further processing and the image displayed exactly as the source.

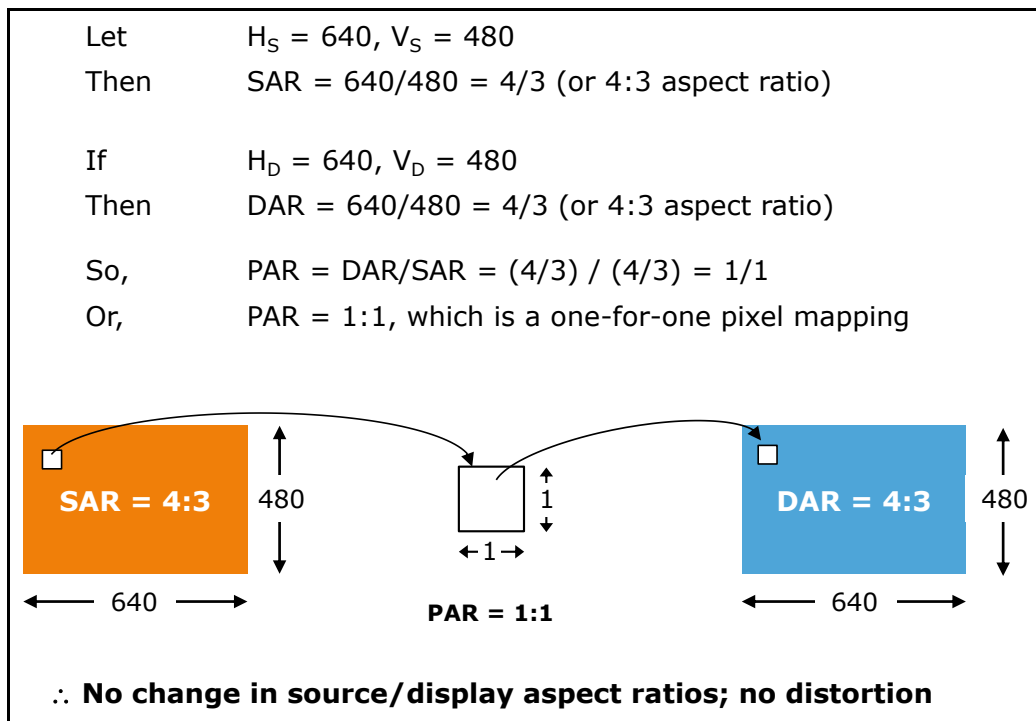


Figure 8: Ideal case of $SAR = DAR$ resulting in $PAR = 1:1$

Figure 9 is an example where the ratio of DAR to SAR is not 1:1 but 8:9. This results in a pixel aspect ratio distortion when the original image pixels have a 1:1 or square ratio.

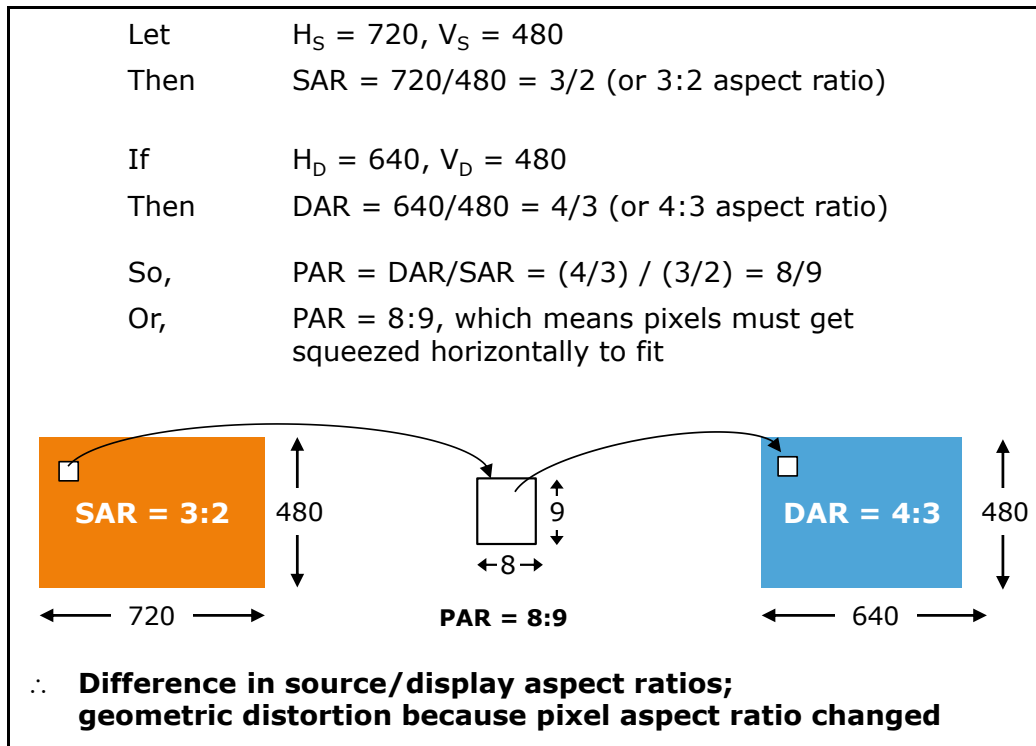


Figure 9: Distortion of pixels

In summary, it is important to preserve the pixel aspect ratio of the original image when applying cropping and scaling. Realizing a receiver display hardware may change the pixel aspect ratio in presentation helps explain why image features appear geometrically incorrect. The various aspect ratio metrics discussed aid in understanding how an image pixel formed by a sensor is displayed by a receiver. Display electronics may change the pixel aspect ratio from square (1:1) to non-square pixels, which changes the geometric shape and features in an image. Such “distortions” can affect algorithms which rely on the accurate measurement of objects within an image.

Appendix B: Image Scaling - Informative

Note: The following is an introduction to the causes and resulting artifacts which may occur when scaling an image.

Image scaling is a signal processing operation which changes an image's size from one format to another; for example, 1280x720 pixels to 640x360 pixels. An image with many pixels does not necessarily imply a higher fidelity image over one with fewer pixels. For instance, if you focus a camera to produce a high fidelity, sharp image, and then take the same picture with the lens of the camera slightly defocused both images will be the same size yet have a very different look; one is sharp and one is blurred. Obviously, the equivalent size of the images did not translate into the desired fidelity. So, image size does not necessarily mean better image quality. What is more important is what information the pixels convey.

Images are comprised of many different frequencies much like those forming a piece of music. However, whereas music is a one-dimensional temporal signal, an image is a two-dimensional spatial signal with horizontal (across a scan line) and vertical (top to bottom) frequency components. Video made from a sequence of images in time adds yet a third dimension of frequency (temporal frame rate). The combination of horizontal, vertical, and temporal frequency components constituting a video signal is termed the spatio-temporal frequency of a video signal.

To simplify the discussion of frequency as related to the number of pixels consider the horizontal dimension of an image only. In typical imagery neighboring pixels tend to have some relation to one another. However, each pixel along a horizontal line can take on a value independent of its neighboring pixels. The maximum change possible from pixel to pixel occurs when sequential pixels' transition from full-on to full-off or zero intensity (black) to 100% intensity (white), or vice versa. The intensity transition in adjacent pixels constitutes one complete cycle -- black-to-white or white-to-black, for example. Conversely, when sequential pixels remain the same value (all pixels are one shade of gray, for example) there is no change, and thus, no frequency change as well; this is defined as zero frequency. Thus, across a line neighboring pixels can vary between some maximum frequency and zero frequency. Horizontal frequency is specified as several of these cycles per picture width (c/pw).

Similarly, the same holds true for vertical pixels within a column of an image. Vertically, frequencies are specified as cycles per picture height (c/ph). In the temporal domain, the maximum frequency is governed by the frame rate, and this is expressed in frames per second, or Hertz.

In the case of the in-focus picture example above, the pixels within the image will exhibit significant change with respect to one another, whereas the defocused picture will have much less change amongst neighboring pixels. A camera's lens acts as a two-dimensional filter, which can smear the received light from the scene onto groups of pixels on the image sensor. In effect, this is equivalent to averaging a neighborhood of pixels and assigning a near constant value to them all.

To gain an appreciation for artifacts image scaling can cause consider what would happen in the example above if each successive pixel across a horizontal line changes from zero to 100% intensity? If this were done for every scan line, the image would look like a series of vertical

stripes each one pixel wide. What would happen if the image is then scaled by one half horizontally, where every other pixel is eliminated? If the eliminated pixels are the zero intensity ones the resulting image would be all white, while if the eliminated pixels are the 100% intensity ones the resulting image would be all black. Obviously, the final scaled image does not resemble the original image. This artifact is called aliasing; named because the resulting frequencies in the signal are completely of a different nature than what they were originally.

An example of aliasing is shown in Figure 10 below.

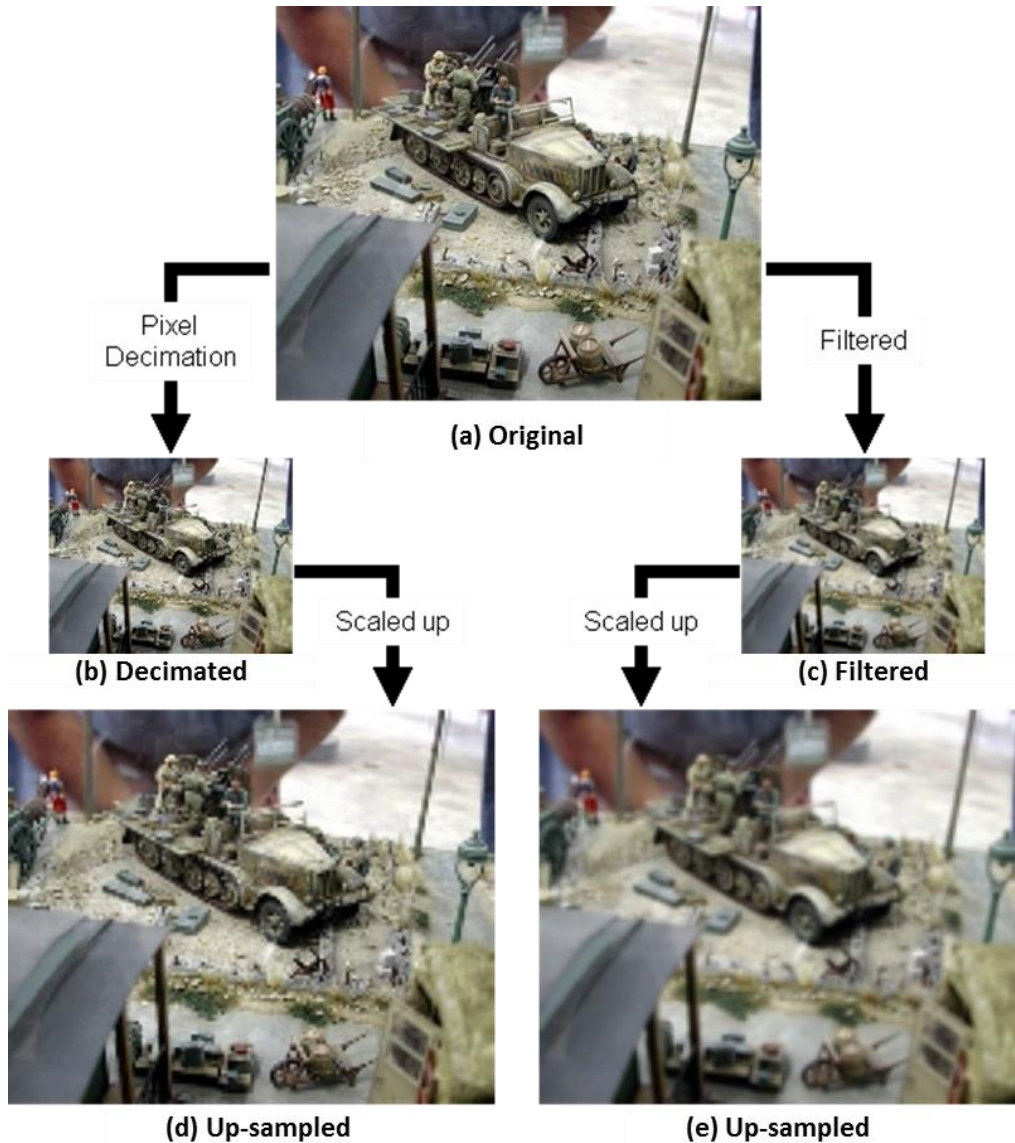


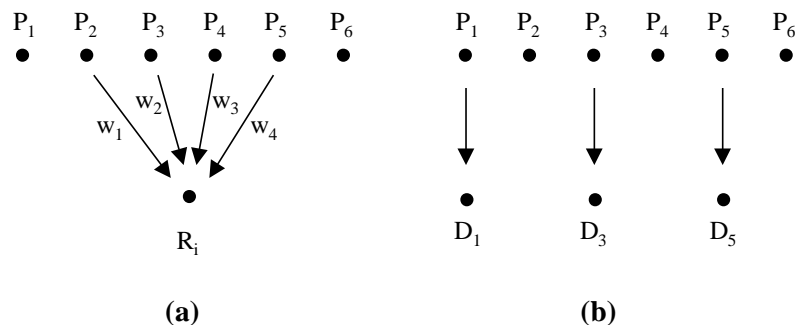
Figure 10: Direct Down-sampling and Filter/Down-sample

The original image in Figure 10(a) is scaled by one half in each dimension using pixel decimation (elimination) in Figure 10(b), and filtering followed by decimation in Figure 10(c). To emphasize the artifacts induced by both techniques, the images are shown up-scaled by two in

Figure 10(d) and Figure 10(e). Although the filtered image appears less sharp, it has far fewer jagged edges and artifacts likely to impact compression negatively.

Filters are signal processing operations used to control the frequencies within a signal, so functions like scaling do not distort the information carried by the original signal. A two-dimensional (2D) filter can remove those spatial frequencies which cannot be supported by the remaining pixels of a scaled image. A 2D low-pass filter, which acts as a defocused lens, is essentially an integrator performing a weighted average of pixels within sub-areas of an image. This integration helps to limit aliasing artifacts. How the integration is done is critical in preserving as much of the image frequency content as possible for a target image size. Some methods of integration can introduce excessive blur or excessive aliasing – both undesirable. Blur reduces image feature visibility, while aliasing produces false information and reduces coding efficiency.

The number of pixels over which a 2D filter operates may be as few as 2x2 (two pixels horizontally by two pixels vertically), which is simply averaging of the four pixels to produce a new one. Such small extent filters are computationally efficient but do a poor job in general. 2D filters which do a better job retain as much image fidelity as possible, and typically include many more neighboring pixels to determine each new scaled output pixel. Figure 11(a) shows a collection of weighted pixels P_k in the horizontal direction which sum to a new output value R_i , while Figure 11(b) shows a direct scaling by two without any filtering. The weights $[w_1-w_4]$ are numerical values multiplied by corresponding pixels with the results added to form a new output pixel. For example, in Figure 11(a) the output pixel $R_i = w_1 \cdot P_2 + w_2 \cdot P_3 + w_3 \cdot P_4 + w_4 \cdot P_5$.



**Figure 11: (a) Input pixels P_k ; filter taps w_1 - w_4 and filtered output pixel R_i
(b) Direct scaling by two**

Alternate pixels are eliminated in direct scaling by two. In this case, the contributions from pixels P_2, P_4 , etc. are completely ignored along with any valuable information they carried.

Spatio-Temporal Frequency

Motion Imagery is a three-dimensional signal with spatial frequencies limited by the lens, the sensor's spatial pixel density, and temporal frequencies limited by the temporal update rate. This collection of 3D frequencies constitutes the spatio-temporal spectrum of a video signal. Scaling in the temporal domain, such as changing from 60 frames-per-second to 30 frames-per-second, is usually accomplished by directly dropping frames rather than applying a filter first. The focus here, therefore, is filtering as applied in the 2D spatial horizontal and vertical dimensions.

When viewed from the frequency perspective, the image will contain horizontal frequencies extending from zero frequency to some maximum frequency limited by the number of horizontal pixels, and likewise, vertical frequencies extending from zero frequency to some maximum frequency limited by the number of vertical pixels. The frequency domain is best understood using a spectrum plot as shown in Figure 12. The amplitudes of the individual component frequencies are suppressed in this figure but would otherwise extend directly outward orthogonal to the page with varying amplitudes.

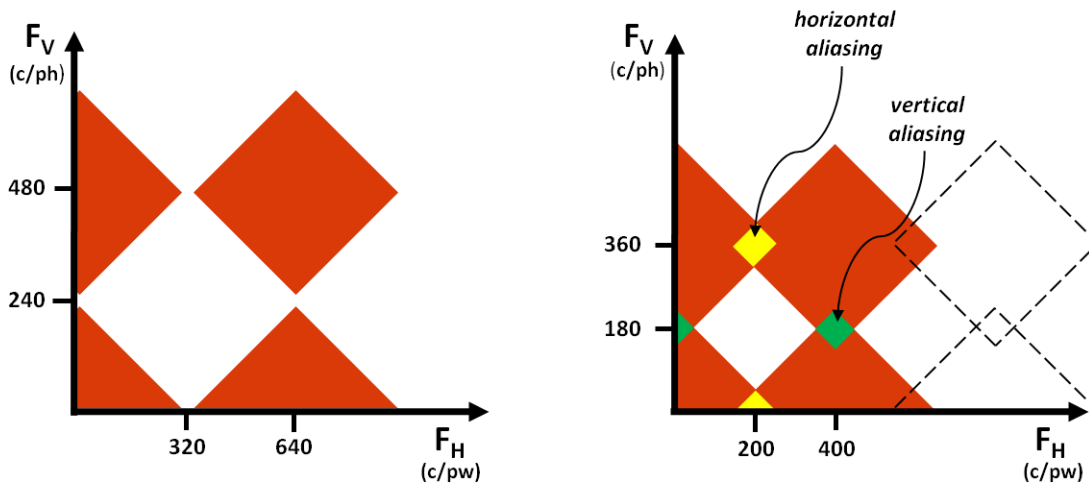


Figure 12: (a) HV Spectrum for a 640x480 image; (b) and re-sampled at 400x360

The value in portraying an image in the frequency domain is the ability to identify potential issues when applying a signal processing operation, such as image scaling. In Figure 12(a), frequencies extend from zero (the origin) to 320 cycles-per-picture width (horizontal frequencies) and 240 cycles-per-picture-height (vertical frequencies). The amplitudes of the frequencies within this quarter triangle depend on the strength of each frequency in the image.

Sampling theory dictates the maximum frequency be no more than one-half the sampling frequency. The sampling frequency for an image is fixed by the number of pixels, and since one cycle represents two pixels the maximum frequency is limited to half the number of pixels in each dimension. A 640x480 image will thus have frequencies no greater than 320 c/pw and 240 c/ph. Most video imagery is limited in spatial frequency extent by the circular aperture of the lens, and so the frequency spectrum is rather symmetrical about the zero-frequency point.

Sampling theory indicates the frequency spectrum signal repeats at multiples of the sampling frequency. A digital image spectrum repeats itself at intervals equal to the picture width and picture height – its sampling frequency. For example, the horizontal spectrum of a 640-pixel image will repeat at intervals of 640 c/pw. The vertical spectrum will repeat at intervals of 480 c/ph for 480 pixels.

If the horizontal, vertical, or temporal sample intervals are too close to one another because of scaling, or reducing the temporal rate, then these repeat spectra will overlap causing image artifacts. This interference produces cross-modulation frequencies which manifest themselves as aliasing (Figure 12(b)) and flicker. On the other hand, if an image is overly filtered, the image may become blurred because too many higher frequencies are attenuated. Scaling an image to a smaller size will re-position the repeating frequency spectrum's closer to one another reflective of the effective sampling rate being lowered. A filter will limit the frequencies in an image in an orientation, so the image can be scaled with minimal artifacts.

Rules of Thumb

Scaling an image will cause artifacts when the resulting pixels can no longer support frequencies contained within the image. The number and values of the filter weights determine the final quality of the scaled image. For good quality scaling between 100-50% (where 100% is the original image size and 50% is half the size in both the horizontal or vertical directions) five filter weights in the direction of scaling is sufficient; nine filter weights are sufficient for scaling 50-25%. For drastic reductions of 25-12.5% 17 filter weights may be required.

These rules of thumb are not required for manufacturers to follow. They are only included for guidance. It is to be appreciated vendors will provide their own value-added solutions.